Updated SNP-wise Mean model in version 1.08

In version 1.08 of MAGMA, the SNP-wise Mean gene analysis model has been updated to address an issue with some inflation of error rates particularly in larger (in number of SNPs) genes that was discovered and outlined in Yurko et al. (2020)[1].

The original implementation[2] used the test statistic $T = -2\sum_j^K \log p_j$, with K the number of SNPs and p_j the marginal p-value for SNP j. Since the sampling distribution does not have a closed form expression, the distribution was approximated with a scaled χ^2-distribution following[3], adding an additional post-hoc transformation of the p-values. Although this yielded well-controlled type 1 error rates in the simulations conducted for the original MAGMA paper[2], the recent simulation study by Yurko et al. (2020)[1] found this no longer to be the case.

The likely cause of the discrepancy is the considerable growth in the SNP density of GWAS data, resulting in much higher LD between SNPs included in a gene on average. This decreases the accuracy of the scaled χ^2-distribution as an approximation of the sampling distribution of T, an effect that would indeed become more pronounced for larger genes.

Updated model

To resolve this issue, the SNP-wise Mean model was changed to use an alternative test statistic $T^* = \sum_j^K Z_j^2 = Z^T Z$, with $Z_j = \Phi(p_j)$ and Φ the cumulative normal distribution function; that is, T^* is the sum over squared SNP Z-statistics. Jointly, for the vector Z we can assume $Z \sim \mathrm{MVN}(0, S)$, where S is the correlation matrix of the SNP genotypes.

Given the eigendecomposition $S = Q\Lambda Q^T$, we rewrite $Z = Q\Lambda^{0.5}D$ for a random variable $D \sim \mathrm{MVN}(0, I_K)$. It follows that $T^* = Z^T Z = D^T\Lambda^{0.5}Q^T Q\Lambda^{0.5}D = D^T\Lambda D = \sum_j^K \lambda_j D_j^2$, with λ_j the jth eigenvalue and $D_j^2 \sim \chi_1^2$. That is, the distribution of T^* is equal to a mixture distribution of independent χ_1^2 random variables.

Imhof[4] provides a method of expressing the gene p-value $P = Pr(T^* \geq T^*_{obs})$ as an integral that can be evaluted numerically, which is the approach implemented for the updated SNP-wise Mean model (using Gauss-Kronrod quadrature for the integration[5]). Whereas previously for T an approximate distribution was evaluated, in this Imhof approach for T^* we directly evaluate the sampling distribution itself. Although this removes the problem that arose with the previous implementation, it is possible for the numerical integration to fail for some genes. We have therefore additionally implemented a fallback procedure that is used in case the numerical integration fails.

This fallback procedure generates empirical p-values, using an optimized simulation process to generate draws from $\sum_j^K \lambda_j D_j^2$. The number of simulations used is determined adaptively to reduce computing time, using up to a maximum of one billion simulations. Empirical p-values that are still zero afterwards are set to 0.5/1e9 = 5e-10. Though the actual p-value in such cases is likely to be significantly lower, running additional simulations beyond this becomes computationally unfeasible. Moreover, any multiple-testing corrected significance threshold that is likely to be used will generally much higher than this lower bound, and precision of the p-values around that threshold will be sufficient to reliably determine significance.

It should be noted that the change in test statistic from T to T^* also represents a slight change in the statistical behavior, aside from the practical aspects of evaluating their respective sampling distributions. Under the gene null hypothesis of no association Z_j^2 has a χ_1^2 distribution, whereas $-2\log p_j$ instead has a χ_2^2 distribution. As such the relative contribution of each SNP to T^* will be somewhat different to what

it was in the original statistic T. This means that there is some redistribution of power across different scenarios of association. In practice the differences are small to negligible, particularly for more strongly associated genes. Similarly, although the change in test statistic also results in some changes to NPARAM values in the MAGMA output, these changes again are minor.

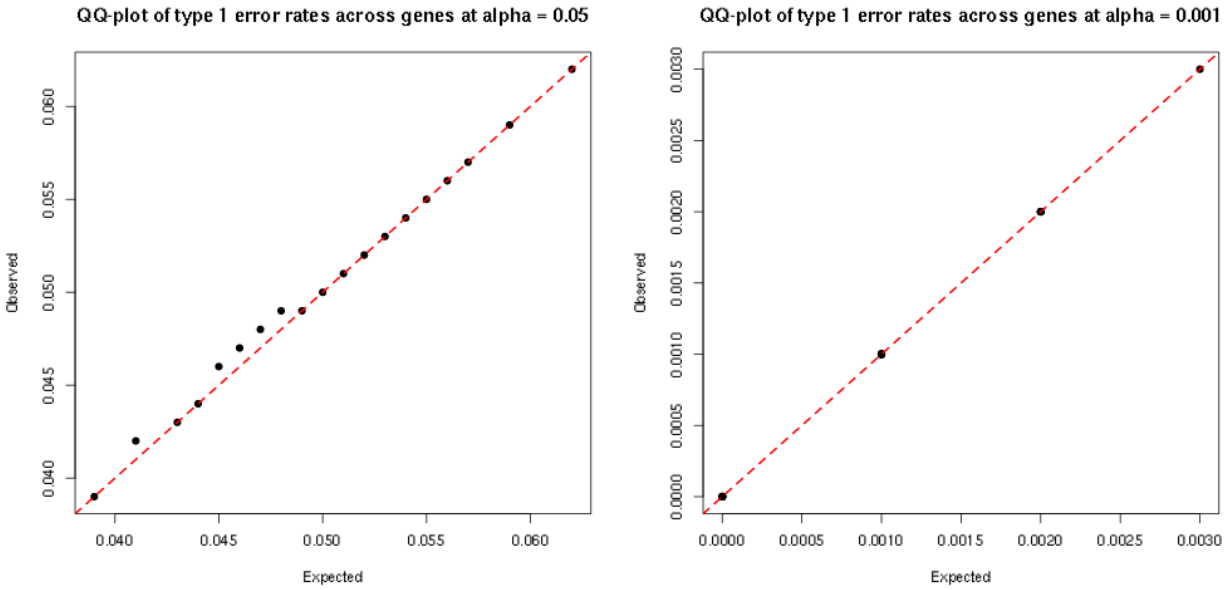
Type 1 error simulations

We performed a simulation study to evaluate the type 1 error rates of the updated model. This was based on the European cohort of the 1,000 Genomes data and the NCBI 37.3 gene definitions that can be found on the MAGMA site.

The genes were first annotated to the data, resulting in a total of 19,248 annotated genes containing 9,220,220 unique SNPs, and 485.4 SNPs per gene on average. We duplicated individuals in the 1,000 Genomes data to attain a sample size of 5,000, and generated 1,000 null phenotypes drawn from a standard normal distribution, each of which was analyzed using the updated SNP-wise Mean model.

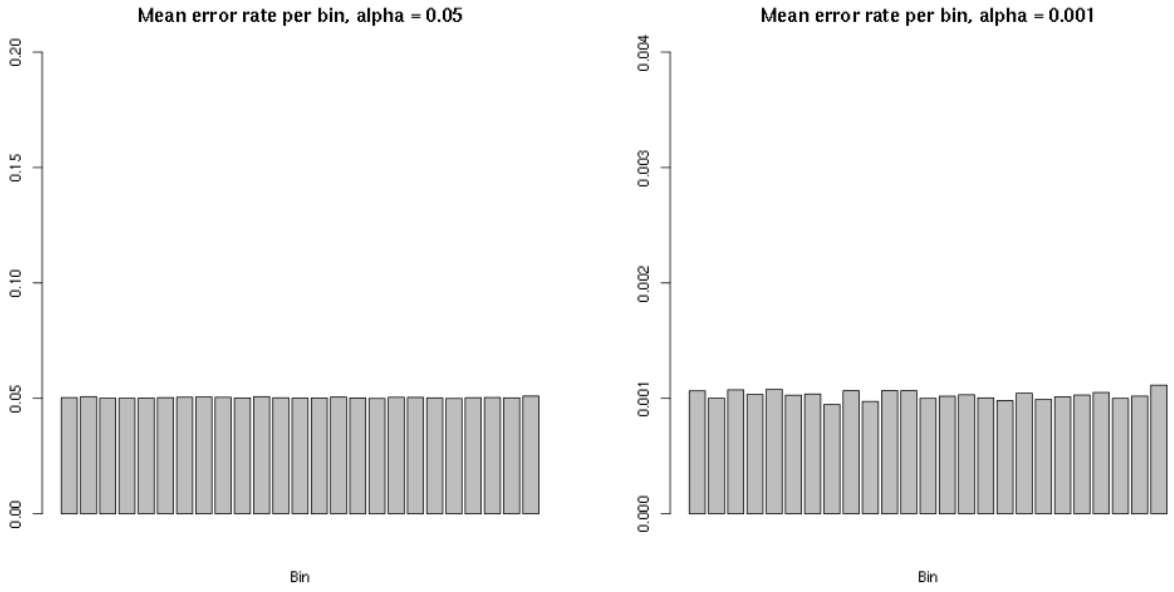
For the output, we first estimated Bonferroni family-wise error rate FWER at $\alpha = 0.05$ as the proportion of the 1,000 simulations for which at least one gene had a p-value below 0.05/19,248. This estimate came to 0.046, showing the FWER to be well-controlled.

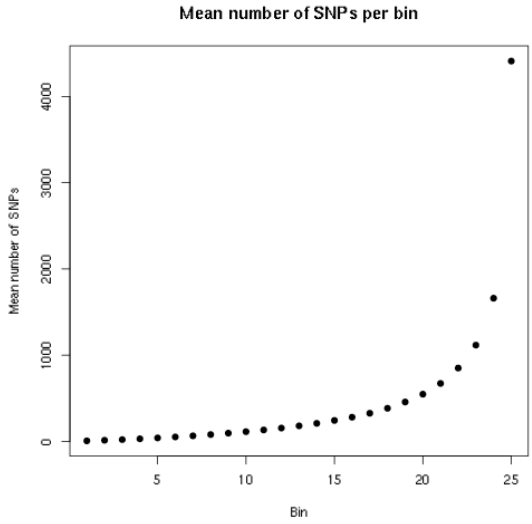
We then computed the type 1 error rate per individual gene at α's of 5% and 0.1% , plotting the quantiles of the error rates across the genes against the expected quantiles (5^{th} to 95^{th} percentiles in steps of 5). These were obtained by taking the quantiles of a binomial distribution with 1,000 trials and a success probability of 5% or 0.1%, and dividing these by 1,000. These are shown in the figure below, and as can be seen this distribution is again consistent with the error rates being well-controlled (the reason only four points are shown for the 0.1% alpha is that because of the low success probability many of the quantiles overlap.



Finally, to determine whether there was any relation between type 1 error rate and gene size, we sorted the genes in ascending order of size, subdivided them in equally sized bins (either 769 or 770 genes per bin), and then computed the average type 1 error rate at both the 5% and 0.1% α's for the genes per bin. Bar plots of these are shown below. As can be seen, beyond some random fluctuation due to simulation noise, the mean type 1 error rate is essentially constant across the bins. For additional reference, a plot of the mean number of SNPs per bin has also been included below.

We also fitted a linear regression model with the gene error rates as outcome, using the gene size S (the NSNPS column in MAGMA output), gene 'density' D (the NPARAMS column), and their ratio $R = D/S$. In total, eight predictors were used: S, D and R themselves, the squared values of all three, and the log values of S and D. The adjusted R^2 values for these models were 0.000235 for the 5% error rate and 0.0012 for the 0.1% error rate, further confirming that the gene type 1 error rates are independent on the size and level of LD of a gene.





References

1. Yurko, R., Roeder, K., Devlin, B. & G'sell, M. H-MAGMA, inheriting a shaky statistical foundation, yields excess false positives. 1–14 (2020) doi:10.1101/2020.08.20.260224.
2. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, 1–19 (2015).
3. Brown, M. B. A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics* **31**, 987–992 (1975).
4. P.J., I. Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**, 419 (1961).
5. Laurie, D. P. Calculation of Gauss-Kronrod quadrature rules. *Math. Comput.* **66**, 1133–1146 (1997).